METHOD FOR GENE IDENTIFICATION SIGNATURE (GIS) ANALYSIS

FIELD OF THE INVENTION

The present invention relates generally to the field of gene and transcript expression and specifically to a method for the serial analysis of a large number of transcripts by identification of a gene signature (GIS) corresponding to defined regions within a transcript.

BACKGROUND OF THE INVENTION

One of the most important goals of the human genome project is to provide complete lists of genes for the genomes of human and model organisms. Complete genome annotation of genes relies on comprehensive transcriptome analysis by experimental and computational approaches. *Ab initio* predictions of genes must be validated by experimental data. An ideal solution is to clone all full-length transcripts and completely sequence them. This approach has gained recognition recently (Strausberg, R.L., et al., 1999, *Science*, 286: 455-457) and progress has been made (Jongeneel C.V., et al., 2003, Proc Natl Acad Sci U S A. 100, 4702-4705). However, due to the complexity and immense volume of transcripts expressed in the various developmental stages of an organism's life cycle, complete sequencing analysis of all different transcriptomes still remains unrealistic.

To get around such a dilemma, a cDNA tagging strategy that obtains partial sequences that represent full transcripts has been developed and widely applied in determining genes and characterizing transcriptomes in the past decade.

In the expressed sequence tag (EST) approach, cDNA clones are sequenced from 5' and/or 3' nds (Adams, M., et al., 1991, *Science*, 252, 1651-1656). Each EST sequence read would generate on average a 500bp tag per transcript. The number of same or overlapping ESTs would manifest the relative level of gene expression activity. Though EST is effective in identifying genes, it is prohibitively expensive to tag every transcript in a transcriptome. In practice, sequencing usually ceases after 10,000 or less ESTs are obtained from a cDNA library where millions of transcripts might be cloned.

To increase the efficiency in sequencing and counting large numbers of transcripts, Serial Analysis of Gene Expression (SAGE) ((Velculescu, V. E., et al., 1995, *Science*, 270, 484–487; Saha S, et al., 2002, *Nature Biotechnology*, 20, 508-12; US6,498,013; US6,383,743) and the recent Massively Parallel Signature Sequencing (MPSS) technique (Mao C., et al., 2000, *Proc Natl Acad Sci USA*, 97, 1665–1670; Brenner S, et al., 2000, *Nature Biotechnology*, 18, 630–634) were developed based on the fact that a short signature sequence (14-20bp) of a transcript can be sufficiently specific to represent that gene.

Experimentally, short tags can be extracted from cDNA one tag per transcript. Such short tags can be efficiently sequenced either by a concatenation tactic (as for SAGE) or by a

hybridization-based methodology for MPSS. For example, in SAGE, multiple tags are concatenated into long DNA fragments and cloned for sequencing. Each SAGE sequence readout can usually reveal 20-30 SAGE tags. A modest SAGE sequencing effort of less than 10,000 reads will have significant coverage of a transcriptome. Transcript abundance is measured by simply counting the numerical frequency of the SAGE tags.

With the availability of many assembled genome sequences in public databases, the use of a short tag strategy for transcriptome characterization is becoming popular (Jongeneel et al., 2003, *Proc. Natl. Acad. Sci. USA* 100: 4702-4705). In theory, short DNA tags of about 20bp can be specifically mapped to a single location within a complex mammalian genome and uniquely represent a transcript in the content of whole transcriptome. However, in reality, there still exist a large number of "ambiguous" SAGE tags (14-21bp) and MPSS tags (17bp) that have multiple locations in a genome, and may be shared by many genes. Limited by the availability of type II restriction enzymes that can cut longer than 21bp, the SAGE method currently can not generate any longer tags to improve specificity.

Further, SAGE and MPSS methods only produce a single signature per transcript in the middle of the gene. In view of the "internal" nature of the tag in a transcript, these methods provide only limited tag information.

Therefore, despite their usefulness in sequencing efficiency, the utility of methods such as SAGE or MPSS is severely undermined by their lack of specificity and consequent inconclusiveness.

There is a need in the art for more efficient methods which retain the sequencing efficiency and at the same time improve the use of the tagging strategy for transcriptome characterization and facilitate the annotation of genomes.

SUMMARY OF THE INVENTION

The present invention solves the problems mentioned above by providing two tags (a ditag) per nucleic acid molecule, therefore increasing the specificity of the tags to represent a nucleic acid molecule (for example a gene).The two tags are extracted from the 5' and 3' ends of the same nucleic acid molecule, and therefore ditags are more informative to reflect the structure of the nucleic acid molecules. Critically, the invention provides a method to link the 5' and 3' tags of the same nucleic acid molecule into a single ditag unit. Therefore, the pairs of 5' and 3' tags that represent the nucleic acid molecule can be easily recognized by simple sequencing analysis. The invention can be used for the identification of new genes, for the measure of transcript abundance in transcriptomes, for the annotation of genome sequences and at the same time enhancing sequencing efficiency.

In particular, the invention provides an isolated oligonucleotide comprising at least one ditag, wherein the ditag comprises two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a nucleic acid molecule.

The oligonucleotide of the invention, further comprises at least two adapters flanking the ditag, wherein each adapter comprises at least one restriction site. In particular, each adapter comprises at least: a first restriction site proximal to the tag which is an asymmetric recognition site (for example, a homing endonuclease recognition site, or a type II recognition site) and at least a second restriction site. The second or further restriction site may be any restriction site known in the art may be used. For example, BamHI. Also, any asymmetric restriction site different from the first restriction site may be used. The recognition site for this enzyme however must be absent from the vector backbone after insertion of the ditag

The nucleic acid molecule may be the full-length sequence of a gene or a fragment thereof. For example, RNA, mRNA, genomic DNA, full-length cDNA or cDNA.

The ditag may vary in nucleotide number. According to one embodiment, it is obtained by splicing the 5' terminus and the 3' terminus of a nucleic acid molecule in presence of at least one restriction enzyme and the size of the sequence tags is determined by the restriction

enzyme used. Accordingly, the number of nucleotides of the ditag can vary according to the restriction enzyme used.

When MmeI is used, this enzyme recognizes a sequence inside each of the two adapters that flank the nucleic acid molecule which one intends to reduce, but cuts inside the nucleic acid molecule forming a tag comprising 19-21 nucleotides. Two such tags may be additionally processed by blunting and ligation to form a ditag comprising 34-38 nucleotides. The ditag is hence obtained by splicing together the 5' terminus and the 3' terminus of the same nucleic acid molecule.

The ditag of the invention can be of any size, preferably 12-60 bp.

The oligonucleotide may comprise a concatemer of ditags, for example 1 to 1000 ditags.

The invention also provides a vector comprising the oligonucleotide of the invention. In particular, the vector comprises at least a nucleic acid molecule and at least two adapters flanking the nucleic acid molecule, wherein each adapter comprises at least: a first restriction site which is a asymmetric restriction site (asymmetric restriction site is, for example, a homing endonuclease recognition site, or a type II recognition site) and at least a second restriction site (for example Bam HI), and the backbone of the vector does not comprise the asymmetric restriction site and the second or further restriction site. A preferable, asymmetric restriction site is the type II restriction site MmeI.

The invention also provides a vector having the sequence indicated in SEQ ID NO:18.

The invention further provides a cDNA library, wherein every cDNA clone comprises the at least one oligonucleotide of the invention.

According to another aspect, the invention also provides a method for preparing at least one oligonucleotide comprising at least one ditag comprising:

producing at least one nucleic acid molecule;

isolating the 5' terminus and the 3' terminus of the nucleic acid molecule or fragment thereof; and

linking the 5' terminus and 3' terminus to create the at least one ditag.

In particular, it is provided a method for preparing at least one oligonucletide comprising at least one ditag comprising:

producing at least one nucleic acid molecule flanked by two adapters;

isolating the 5' terminus and the 3' terminus of the nucleic acid molecule; and

linking the 5' terminus and 3' terminus to create the at least one oligonucleotide comprising at least one ditag flanked by the two adapters.

The nucleic acid molecule desired to be reduced in form of a ditag may be a full nucleic acid molecule or a portion inside the nucleic acid molecule.

The nucleic acid molecule may correspond to the full-length of a gene or fragment thereof

The method may further comprise the step of determining the nucleotide sequence of the at least one ditag to detect gene expression.

According to a further aspect, the method of the invention may further comprise the steps of:

determining the sequence of the at least one ditag; and

comparing the ditag nucleotide sequence to a database comprising genomic sequences whereby matching 5' and 3' termini sequences are identified.

According to a particular embodiment, the invention provides a method comprising :

producing at least one nucleic acid molecule, preferably a full-length cDNA, flanked by two adapters, wherein each adapter comprises at least one restriction site; splicing the 5' terminus and the 3' terminus of the nucleic acid molecule to produce at least one ditag by adding at least one restriction enzyme recognizing the recognition sites.

Preferably, each adapter comprises at least: a first restriction site which is an asymmetric restriction site and a second restriction site.

As restriction enzyme, any useful enzyme can be used. For example, a restriction enzyme recognizing two asymmetric recognition sites.

Asymmetric recognition site can be: i) homing endonuclease asymmetric recognition site sequences or ii) restriction endonuclease asymmetric cleavage sites sequences recognizable by type II restriction enzymes.

According to a particular embodiment, the splicing step is carried out by using MmeI (together with T4 DNA polymerase and T4 DNA ligase) and the ditag of 34-38 nucleotides, flanked by two adapters, is produced..

According to a further aspect, the ditag of any embodiment of the invention can be linked to other ditags to produce concatemers of ditag. For example, 1 to 1000 ditags.

According to another further aspect, it is provided a method for genome mapping, comprising:

preparing at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a nucleic acid molecule, the nucleic acid molecule corresponding to the full-length of a gene or fragment thereof;

mapping each of the two tags of the at least one ditag on the genome; and

defining the structural region of the corresponding gene on the genome map.

According to a still another aspect, the invention provides a method of gene discovery comprising:

preparing at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a nucleic acid molecule, the nucleic acid molecule corresponding to the full-length of a gene or fragment thereof;

comparing the obtained at least one ditag with a genome map and/or a gene database;

if the 5' and 3' termini tags of a ditag are matched to the genome sequence but not in known gene databases, then the detected ditags may represent new genes in the given genomes.

Such ditags can directly guide the process of recovering the full-length nucleic acid molecule corresponding to the newly identified genes.

It is also an aspect of the invention a method for recovering the full-length cDNA of new and/ or other interesting genes comprising:

preparing, from a full-length cDNA library, at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a full-length cDNA library;

sequencing the obtained oligonucletide ditag, preferably a large number of the obtained ditags;

determining the ditag of interest (for example, based on biological aspects); and

recovering the full-length cDNA corresponding to the ditag of interest from the parental full-length cDNA library.

Further, the invention also provides a method for quantifying the transcriptional activity of a gene comprising:

preparing, from a full-length cDNA library, at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a full-length cDNA;

sequencing the obtained oligonucleotide ditag, preferably a large number of the obtained ditags;

determining the frequency of the sequenced ditag which corresponds to the transcriptional activity of the gene.

BRIEF DESCRIPTION OF THE FIGURES

FIGURE 1 shows the GIS analysis experimental workflow (bacterial transformation approach). In the figure, the letters N, B, M, S either in capital or small letters denotes the recognition sites for the restriction enzymes Not I, Bam HI, Mme I and Sal I, respectively. The text "Me" represents methylation of the newly-synthesized first-strand cDNA.

FIGURE 2 shows the GIS analysis experimental workflow (PCR-based approach). In the figure, the letters N, B, M, S either in capital or small letters denotes the recognition sites for the restriction enzymes Not I, Bam HI, Mme I and Sal I, respectively. The text "Me" represents methylation of the newly-synthesized first-strand cDNA.

FIGURE 3 shows the GIS application of mapping transcriptome to genome.

FIGURE 4 is an electrophoresis gel showing MmeI digestion of a mix of original full-length cDNA clones. Lane 1: original supercoiled plasmid preparation. Lane 2: 1kb DNA ladder. Lane 3: MmeI digestion products. The arrowhead shows the position of all the linearized tagged-plasmids.

FIGURE 5 is an electrophoresis gel related to the preparation of GIS ditags. The plasmid DNA of GIS ditag library is digested with BamHI. The 50bp ditag fragments are separated and purified from the vector using a 10% polyacrylamide gel. Lane 1: DNA size markers. Lane 2-8: formation of 50 bp GIS ditags.

FIGURE 6 is an electrophoresis gel related to the preparation of GIS ditags by PCR. The ditag-containing PCR fragments generated from the GIS full-length cDNA library are digested by BamHI. The 50bp ditag fragments are separated and purified from adaptor arms in 10% polyacrylamide gel. Lane 1: DNA size markers. Lane 2-15: large scale preparation of 50 bp GIS ditags.

FIGURE 7 shows the pGIS1 vector construct.

FIGURE 8 shows the commercial pZErO-1 vector construct (Invitrogen) The positions of the various sequencing/ PCR primer binding sites (PMR003, PMR004, PMR011 and PMR012) are shown.

FIGURE 9 shows a typical example of the QC (quality check) performed on multiple clones from the GIS library using PCR. Lane 1: pZErO-1 vector as negative control. M: 1kb+ DNA ladder. Lanes 2-25: randomly-picked clones.

FIGURE 10 shows the double strand nucleotide sequence of pGIS1. The region between the restriction sites Not I and Sal I is the stuffer fragment that is removed during cloning. It is highlighted in bold and italic type. The single strand nucleotide sequence is also reported as SEQ ID NO:18. The region representing the stuffer fragment is between nucleotide 15 to 704 (both nucleotides included).

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a Gene Identification Signatures (GIS) and a GIS analysis method: useful, for example, for the rapid analysis of numerous transcripts in order to identify the overall pattern of transcript expression (transcriptome), for the selection and/or construction of cDNA and full-length cDNAs, tag sequencing, gene discovery, genome mapping and annotation. In general, the GIS and GIS analysis method according to the invention greatly facilitates the collection of gene information by experimental approach.

For the purpose of the present application, GIS means a ditag (also indicated as GIS ditag) or an oligonucleotide comprising at least one ditag, wherein the ditag comprises the 5' terminus (or end region) and the 3' terminus (or end region) of a nucleic acid molecule, which it is desired to reduce, "shrink" or represent.

The ditag is shorter than the original nucleic acid molecule from which it originates or which it represents. Preferably, the ditag must be much shorter than the original nucleic acid molecule. As consequence of the "shrinking", the ditag essentially comprises the 5' end region (also indicated as 5' tag) and 3' end region (also indicated as 3' tag) of the original nucleic acid molecule. Hence, the portion of the original nucleic acid molecule which is between or inside the 5' tag and 3' tag is not included in the ditag. The ditag according to the invention retains the most informative features of the original nucleic acid molecule, viz. the start and the end signatures of the nucleic acid. It is thereby also more specific and accurate than SAGE or MPSS methods in characterizing transcriptomes and defining gene structure by mapping the GIS tags to genome sequences.

Accordingly, the invention provides an isolated oligonucleotide comprising at least one ditag, wherein the ditag comprises two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a nucleic acid molecule or fragment thereof.

The oligonucleotide of the invention may further comprise two adapters flanking the ditag, wherein each adapter comprises at least one restriction site (see Figure 1 and Figure 2). In particular, each adapter comprises at least: a first restriction site which is an asymmetric restriction site and at least a second adjacent restriction site. Therefore, the number of restriction sites present in each adapter may be two or more. Examples of asymmetric

restriction sites are homing endonuclease asymmetric recognition sites, and type II (or class II) recognition sites. A list of possible asymmetric restriction sites and corresponding restriction enzymes recognizing such asymmetric sites is reported below. Example of second and further restriction sites may be for example BamHI. This second restriction site is for the purpose of subsequent isolation of a pool of ditags that can then be ligated together to form concatemers.

The original nucleic acid molecule that one intends to reduce (to shrink) may be any natural, any modified or any synthetic nucleic acid molecule. It can also be of any size. The nucleic acid molecule can be a gene (the full-length of a gene) or a fragment thereof. The nucleic acid may be RNA, mRNA, genomic DNA, full-length cDNA, or cDNA or a fragment thereof.

The ditag can also be fully chemically synthesized by comprising the 5' end and 3' end of a nucleic acid molecule which the ditag intends to represent.

The molecule that one intends to reduce may also be a portion or fragment inside a nucleic acid molecule. Accordingly, it is possible to use restriction enzymes recognizing restrictions sites flanking the region which is intended to be reduced. The desired restriction sites may be placed into the appropriate position during the preparation of the nucleic acid molecule, for example a cDNA or full-length cDNA.

According to a particular aspect, the nucleic acid desired to be reduced is a full-length cDNA. Full-length cDNA can be prepared according to any method known in the art. See for example, the cap-trapper approach, for example Carninci et al., 1996, Genomics, Vol.37, 327-336; US 6,143,528; Edery et al., 1995, Mol. Cell. Biol., Vol.15, No.6, 3363-3371.

Those of skill in the art will know other capture systems, for example, those based on biotin/streptavidin, digoxigenin/anti-digoxigenin for isolation of the full-length cDNAs can be used.

The ditag can be prepared according to any technique known in the art. For example, the original nucleic acid molecule may be cut through any chemical reaction and the obtained 5' and 3' termini ligated to create the ditag.

The nucleic acid molecule which is intended to be reduced, which is preferably prepared comprising two adapters flanking the molecule, may be inserted into a vector. In a particular

realisation, each adapter comprises at least one restriction site, preferably comprises at least a first restriction site comprising an asymmetric restriction site and a second restriction site. Accordingly, in the vector used, it is important that the backbone of the vector does not comprise the restriction site or sites present in the adapters.

Accordingly, a library of nucleic acid molecule (for example, a library of full-length cDNAs) is prepared.

Preferably, the nucleic acid molecule is spliced into a ditag or oligonucleotide comprising a ditag by using restriction enzymes which recognize restriction sites flanking the nucleic acid molecule to be reduced. Accordingly, the recognition sites are placed upstream of the 5' terminus and downstream of the 3' terminus of the nucleic acid molecule or fragment thereof desired to be reduced (preferably into the adapters). Accordingly, the oligonucleotide obtained by splicing comprises two adapters flanking the ditag. Each adapter comprising at least one restriction site. Preferably, comprising at least one first restriction site which is an asymmetric site (for example a type II restriction site, like MmeI) and at least a second restriction site (any known restriction site may be used, for example BamHI.

The 5' tag and 3' tag forming the ditag may have the same or different size. Preferably, they have the same number of nucleotides.

The ditag can be of any size, but needs to be meaningful and advantageous over the size of the parental sequence from which it is derived. The preferred size of a tag or ditag is determined by genome complexity. For a bacterial genome a tag from about 8 bp to about 16 bp may be sufficient whereas for a complex genome like the human genome, a 16-20 bp tag (or in other words a 32-40bp ditag) may be considered. In general, the size of the ditag is from about 12-60 bp.

For the purpose of the present application, the terms 5'-terminus, 5'-end and 5'-tag are equivalent to each other and can be used interchangeably. In the same way, the terms 3'-terminus, 3'-end and 3'-tag are equivalent to each other and can be used interchangeably. In an original nucleic acid molecule or portion inside a nucleic acid molecule that one intends to reduce or represent, each of the 5'-end and 3'-end represents a region or portion most closer to the extremity and most far from the middle region of the molecule.

According to one aspect, the 5'-tag and 3'-tag comprised in the ditag are the regions of the molecule cleaved by a restriction enzyme most closer to the 5'-end and 3'-end, respectively, of the nucleic acid molecule or portion thereof which is intended to be reduced or represented. Accordingly, the size of the ditag can be determined by the restriction enzyme or enzymes used. The invention, therefore, relates to an oligonucleotide comprising at least one ditag, wherein the ditag is obtained by splicing the 5' terminus and the 3' terminus of the nucleic acid molecule in the presence of at least one restriction enzyme, which recognizes the restriction sizes flanking the nucleic acid molecule. Accordingly, the size of the sequence tags is determined by the restriction enzyme used.

When preparing the nucleic acid molecule, for example a full-length cDNA, desired restriction sites flanking the 5'-end and 3'-end of the region which is intended to be reduced or represented are inserted. An example of construction of a full-length cDNA by insertion of desired restriction sites flanking the 5'-end and 3-end is shown in Fig. 1 and Fig.2. A full-length cDNA library is then prepared, following which a GIS ditag library is subsequently prepared.

As an example, a restriction enzyme recognizing an asymmetric restriction site can be used for the purpose of the preparation of the ditag according to the invention. In particular a type II enzyme, for example MmeI.

As an example, asymmetric sites can be introduced. Asymmetric site sequences useful for the purpose of the present invention are: i) two homing endonuclease asymmetric recognition site sequences or ii) restriction endonuclease asymmetric cleavage sites sequences recognizable by type II restriction enzymes.

Homing endonucleases are sold and described by New England Biolabs, Inc.; a description of the asymmetric site sequences is also available in the New England Biolabs Catalog. These homing endonuclease asymmetric recognition site sequences are from 18 to 39 bp. However, in the present invention the recognition site sequences are not limited to those sequences nor to these sizes. Preferably, the restriction homing endonucleases capable of cutting the asymmetric site sequences are selected from the group consisting of : I-CeuI, PI-SceI, PI-PspI and I-SceI. The list mentioned above however is not exhaustive. Other homing endonucleases known in the art and those which may be later discovered are included in the scope of the present invention.

Examples of type II restriction enzymes include:

AarI, AceIII, AloI, BaeI, Bbr7I, BbvI, BbvII, BccI, Bce83I, BceAI, BcefI, BcgI, BciVI, BfiI, BinI, BplI, BsaXI, BscAI, BseMII, BseRI, BsgI, BsmI, BsmAI, BsmFI, Bsp24I, BspCNI, BspMI, BsrI, BsrDI, BstF5I, BtgZI, BtsI, CjeI, CjePI, EciI, Eco31I, Eco57I, Eco57MI, Esp3I, FalI, FauI, FokI, GsuI, HaeIV, HgaI, Hin4I, HphI, HpyAV, Ksp632I, MboII, MlyI, MmeI, MnlI, PleI, PpiI, PsrI, RleAI, SapI, SfaNI, SspD5I, Sth132I, StsI, TaqII, TspDTI, TspGWI, TspRI and Tth111II (the list in the web site of Rebase Enzymes®: http://rebase.neb.com/cgi-bin/outsidelist; see also Szybalski, W., 1985, Gene, 40:169). The list mentioned above however is not exhaustive. Other type II enzymes known in the art and those which may be later discovered are included in the scope of the present invention.

Examples of recognition sites and cleavage sites of several class II restriction enzymes are (into parenthesis are the recognition site and the cleavage site): BbvI (GCAGC 8/12), HgaI (GACGC 5/10), BsmFI (GGGAC 10/14) SfaNI (GCATC 5/9), and Bsp I (ACCTGC 4/8).

The ditag of the invention can conveniently be ligated or joined in order to form concatemers of ditag. Accordingly, the invention relates to an oligonucleotide comprising 1 to 1000 ditags, in particular 1 to 200, more in particular 8 to 20 ditags. When ditags are concatemerized, a higher yield of information is achieved because the oligonucleotide, vector or clone comprises more ditags. Hence, the concatenation of ditags allows an efficient analysis of the nucleic acid molecules, like full-length cDNAs, in a serial manner by sequencing multiple ditags within a single vector or clone.

The oligonucleotide, ditag or concatemers of ditags can be inserted into a vector either before or after concatemerization.

According to one aspect, the oligonucleotide comprising the ditag is amplified. For example, by using PCR or any other known amplification methods. Accordingly, suitable PCR primers corresponding to specific regions inside the vector are used. Such regions flank the oligonucleotide comprising the ditag and adapters. PCR can be performed directly on the ligation (self-circularization) reaction to obtain short (for example 200bp) PCR products (see the PCR approach in Figure 2). These PCR products that contain the required GIS ditags will

then be cut with an enzyme recognizing the at least second restriction site (inside the adapters) to generate the required short cohesive ditags. As restriction enzyme recognizing the second or further restriction site, BamHI can for example be used, and cohesive ditags of 50bp are generated. The advantage of this amplification step is that of generating GIS ditags circumventing the need to produce a GIS ditag library amplification, which can be avoided by not transforming the self-circularized tagged plasmids. The amplified oligonucleotide can then subsequently be excised from the vector (in this example, by digestion with BamHI) and concatenated in long stretches of DNA or RNA for subsequent cloning and sequencing analysis (see Fig.1 and Fig.2).

As a particular aspect, the invention discloses a cDNA library wherein the oligonucleotide comprises at least one ditag, and wherein the ditag comprises 34-38 nucleotides and is obtained by splicing nucleotides from the 5' terminus and nucleotides from the 3' terminus of a full-length cDNA or fragment thereof.

The ditag library according to the invention is representative of the library comprising the original nucleic acid molecules. For example, when the library comprising the nucleic acid molecules is a full-length cDNA library, the ditag library is representative of the full-length ditag library Each ditag clone comprises sufficient information characterizing the specific full-length clone. More important, the ditag of the invention comprises the 5'-end and 3'-end of the original full-length cDNA. Hence, the ditag is representative of the structure of the full-length cDNA.

Accordingly, it is sufficient to sequence and analyze the ditag clones of the ditag library. In case a ditag of interest is found, the corresponding full-length cDNA can be selected and prepared from the full-length cDNA library, for example by PCR or directly from target RNA samples by RT-PCR.

The invention provides a method for the preparation of at least one oligonucleotide comprising at least one ditag comprising:

producing at least one nucleic acid molecule;

isolating the 5' terminus and the 3' terminus of the nucleic acid molecule or fragment thereof;

linking the 5' terminus and 3' terminus to create the at least one ditag.

In particular, the invention provides a method for preparing at least one oligonucleotide comprising at least one ditag comprising:

producing at least one nucleic acid molecule flanked by two adapters;

isolating the 5' terminus and the 3' terminus of the nucleic acid molecule; and

linking the 5' terminus and 3' terminus to create the at least one oligonucleotide comprising at least one ditag flanked by the two adapters.

The method further comprising including the oligonucleotide comprising the at least one ditag flanked by the adapters into a vector.

The nucleic acid molecule which is intended to shrink or represent may be RNA, mRNA, genomic DNA, full-length cDNA, or cDNA.

The nucleic acid molecule may be the full-length sequence of a gene or a fragment thereof.

The method of the invention may further comprise the step of determining the nucleotide sequence of the at least one ditag to detect gene expression.

The method may further comprise the steps of: determining the sequence of the at least one ditag; and comparing the ditag nucleotide sequence to a database comprising genomic sequences whereby matching 5' and 3' termini sequences are identified.

More in particular, the invention relates to a method comprising:

producing at least one nucleic acid molecule, for example a full-length cDNA, flanked by two adapters, wherein each adapter comprises at least one restriction;

splicing the 5' terminus and the 3' terminus of the nucleic acid molecule or fragment thereof to produce at least one ditag by adding at least one restriction enzyme recognizing the recognition sites.

Any recognition site known in the art may be used. Restriction enzyme recognizing at least one recognition site within the nucleic acid molecule and which can be used will be evident to those skilled in the art (see for example, Current Protocols in Molecular Biology, Vol. 2, 1995, Ed. Ausubel, et al., Greene Publish. Assoc. & Wiley Interscience, Unit 3.1.15; New England Biolabs Catalog, 1995).

For example, the two recognition sites may be asymmetric recognition sites:

The asymmetric recognition site are: i) homing endonuclease asymmetric recognition site sequences or ii) restriction endonuclease asymmetric cleavage sites sequences recognizable by type II restriction enzymes.

The type II restriction enzyme is selected from the group consisting of AarI, AceIII, AloI, BaeI, Bbr7I, BbvI, BbvII, BccI, Bce83I, BceAI, BcefI, BcgI, BciVI, BfiI, BinI, BplI, BsaXI, BscAI, BseMII, BseRI, BsgI, BsmI, BsmAI, BsmFI, Bsp24I, BspCNI, BspMI, BsrI, BsrDI, BstF5I, BtgZI, BtsI, CjeI, CjePI, EciI, Eco31I, Eco57I, Eco57MI, Esp3I, FalI, FauI, FokI, GsuI, HaeIV, HgaI, Hin4I, HphI, HpyAV, Ksp632I, MboII, MlyI, MmeI, MnlI, PleI, PpiI, PsrI, RleAI, SapI, SfaNI, SspD5I, Sth132I, StsI, TaqII, TspDTI, TspGWI, TspRI and Tth111II (see the list in the web site of Rebase Enzymes®: http://rebase.neb.com/cgi-bin/outsidelist; see also Szybalski, W., 1985, Gene, 40:169; and). The list mentioned above however is not exhaustive. Other type II enzymes known in the art and those which may be later discovered are included in the scope of the present invention.

The enzyme recognizing the homing endonuclease asymmetric restriction site is selected from the group consisting of : I-CeuI, PI-SceI, PI-PspI and I-SceI. The list mentioned above however is not exhaustive. Other homing endonucleases known in the art and those which may be later discovered are included in the scope of the present invention.

A particularly preferred tagging enzyme, according to the invention, is an enzyme which cleaves 20/18 nucleotides 3' of its recognition site forming 3' overhanging ends, such as MmeI

Artificial restriction endonucleases can also be used. These endonucleases may be prepared by protein engineering. For example, the endonuclease FokI has been engineered by insertions so that it cleaves one nucleotide further away from its recognition site on both strands of the DNA substrates. See Li and Chandrasegaran, Proc. Nat. Acad. Sciences USA 90:2764-8, 1993. Such techniques can be applied to prepare restriction endonucleases with desirable recognition sequences and desirable distances from recognition site to cleavage site.

The method further comprises producing concatemers of ditag. The concatemers may be generally about 1 to 1000 ditags, in particular 1 to 200 ditags, more in particular 8 to 20 ditags. While these are preferred concatemers, it will be apparent that the number of ditags which can be concatenated depends on the length of the individual tags and can be readily determined by those of skilled in the art without undue experimentation. After formation of concatemers, multiple tags may be cloned into a vector for sequence analysis, or ditags or concatemers can be directly sequenced without cloning by methods known to those of skill in the art.

The ditags present in a particular clone can be sequenced by standard methods (see for example, Current Protocols in Molecular Biology, supra, Unit 7) either manually or using automated methods.

As described above, the method comprises introducing the oligonucleotide comprising the at least one ditag in a vector.

With the term vector or recombinant vector it is intended a plasmid, virus or other vehicle known in the art that has been manipulated by insertion or incorporation of the ditag genetic sequences. Such vectors contain a promoter sequence which facilitates the efficient transcription. The vector typically contains an origin of replication, a promoter, as well as specific genes which allow phenotypic selection of the transformed cells. Vectors suitable for use in the present invention include for example, pBlueScript (Stratagene, La Jolla, CA); pBC, pZErO-1 (Invitrogen, Carlsbad, CA)(see Fig.8) and pGEM3z (Promega, Madison, WI) or modified vectors thereof as well as other similar vectors known to those of skill in the art. As a particular realisation, the pGEM3z vector has been modified, and will be referred to as pGIS1 (see also Figures 7 and 10). pGEM vectors have also been disclosed in US 4,766,072, herein incorporated by reference.

For the production of the parental nucleic acid molecule, for example full-length libraries and the GIS ditag libraries, suitable vectors are used. Accordingly, suitable vectors, which are within the scope of the present invention, are those wherein the backbone of the vector does not comprise the same restriction site comprised in the adapters flanking the parental nucleic acid molecule or the ditag, after insertion of the parental nucleic acid molecule. Preferably, the invention provides a vector wherein the vector backbone (other than within the stuffer region that is removed during insertion of the parental nucleic acid molecule) does not

comprise the asymmetric restriction site and the second or further restriction site which are comprised into the adapters. In particular, the vector does not comprise the at least asymmetric II restriction site (for example type II restriction site) and the at least second or further restriction site comprised in the adapters. More preferably, the vector backbone (other than within the stuffer region that is removed during insertion of the parental nucleic acid molecule) does not comprise MmeI and BamHI.

An example of such a vector not comprising MmeI in any region outside of the stuffer is the vector pGIS1 shown in Figure 7 and Figure 10. In pGIS1 the MmeI recognition sites were deleted by mutagenesis. The sequence is shown in Figure 10 and in SEQ ID NO:18. In Figure 10, the stuffer region between the sites Not I and Sal I has been highlighted. The invention also related to the pGIS vector comprising the oligonucleotide according to any embodiment of the invention.

The oligonucleotide(s), ditag(s) or concatemer(s) of the invention may also be ligated into a vector for sequencing purposes.

Vectors in which the ditags are cloned can be transferred into a suitable host cell. Host cells are cells in which a vector can be propagated and its DNA expressed. The term also includes any progeny of the subject host cell. It is understood that all progeny may not be identical to the parental cell since there may be mutations that occur during replication. However, such progeny are included when the term host cell is used. Methods of stable transfer, meaning that the foreign DNA is continuously maintained in the host, are known in the art.

Transformation of a host cell with a vector containing ditag(s) may be carried out by conventional techniques as are well known to those skilled in the art. Where the host is prokaryotic, such as E. coli, competent cells which are capable of DNA uptake can be prepared from cells harvested after exponential growth phase and subsequently treated by the $CaCl_2$ method using procedures well known in the art. Alternatively, $MgCl_2$ or RbCl can be used. Transformation can also be performed by electroporation or other commonly used methods in the art.

An embodiment of this is shown in Fig.1 and Fig.2. According to this embodiment, the method of the invention comprises:

producing at least one nucleic acid molecule comprising a full-length cDNA molecule flanked by two adapters; each adapter comprising MmeI recognition sites and another recognition site, which may be BamHI, flanking the 5' terminus and 3' terminus of the full-length cDNA;

splicing the 5' terminus and the 3' terminus of the full-length cDNA to produce at least one ditag, comprising cleaving the full-length cDNA with MmeI which forms 3' overhanging tag ends, and ligating the two 5' and 3' termini tags to produce the ditag.

As shown in Figure 1 and Figure 2, the use of restriction enzymes may leave 5' and 3' double stranded end comprising a short overhanging end (also referred to as sticky end or cohesive end) consisting of few nucleotides. In particular, by using MmeI, the produced 5' and 3' ends consist each of a 20 bp double strand and two nucleotides as 3' overhanging ends. The two tags may be followed by blunt-ending and intra-molecular self ligation to produce tagged plasmids that contain 18 bp signature sequence as 5' end and another 18 bp signature sequence as 3' end of the parental transcript. However, the number of nucleotides cut by MmeI is variable. Accordingly, the ditag obtained by using MmeI may be of 34-38 bp.

The vector which has been used for the preparation of full-length cDNA library is pGIS1. As mentioned above, pGIS1 does not contain in its backbone MmeI restriction sites, other than within the stuffer region between Not I and Sal I, this stuffer region being subsequently removed during production of the libraries.

The oligonucleotide comprising the ditag flanked by the adapters is cut out form the GIS ditag library and linked to other oligonucleotides comprising ditag and adapters to form concatemers of ditags. The concatemers of ditag are then cloned into a vector for sequencing analysis.

Before cutting the oligonucleotide out from the GIS ditag library, it can be amplified directly from the ligation (self-circularization) reaction mix, for example by PCR using suitable primers. The recovered amplified oligonucleotide comprising ditag and adapters is then linked to other oligonucleotides comprising ditag and adapters to form concatemers of ditags. The concatemers of ditag are then cloned into a vector for sequencing analysis.

The method may further comprise the steps of:

determining the nucleotide sequence of the ditag;

detecting the gene expression;

and/or comparing the determined nucleotide sequence to a database comprising genomic sequences whereby matching 5' and 3' termini sequences are identified.

In particular, the at least one ditag comprises 36 nucleotides and the first and second sequence tags comprise each 18 nucleotides.

As mentioned above, the ditag according to the invention includes the "signature" (consisting of the 5' and 3' ends) of the nucleic acid molecule which is intended to be reduced or represented. Such ditags, preferably cDNA ditags, of a library may be concatenated and sequenced. The paired 5' and 3' signature sequences (tags) of a transcript in a ditag delineate the starting and ending points of transcripts. The ditag can be split up in the two tags during data analysis and mapped head-to-head in a specific region within a reasonable distance on a chromosome of an assembled genome sequence. The genomic DNA sequence in between these two tags is the full structural content of the prospective gene, including exons and introns.

A general description of genome mapping using the ditag of the invention is shown in Fig.3.

A modest sequencing run can generate sufficient data to characterize a transcriptome not only by determining the level of transcript abundance but also by defining the structure of transcripts using the revealed 5'and 3' regions. This results in about over 20-fold more efficient than EST sequencing

Because the tags of the ditag can be matched to any genome, for example to human genomic sequences, PCR and RT-PCR primers can then be designed based on the matching genomic sequence.

Accordingly, a further aspect of the invention relates to a method for genome mapping, comprising:

preparing at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a

nucleic acid molecule, the nucleic acid molecule corresponding to the full-length of a gene or fragment thereof;

mapping each of the two tags of the at least one ditag on the genome; and

defining the structural region of the corresponding gene on the genome map.

Further, it is also an aspect of the invention to provide a method of gene discovery comprising:

preparing at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a nucleic acid molecule, the nucleic acid molecule corresponding to the full-length of a gene or fragment thereof;

comparing the obtained at least one ditag with a genome map and/or a gene database;

detecting matching of the 5' and 3' termini tags on the genome map but detecting no match on one or more of the known gene database;

The method further comprises the step of recovering the full-length nucleic acid molecule corresponding to the newly discovered gene.

The invention also provides a method for recovering full-length cDNA comprising:

preparing, from a full-length cDNA library, at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a full-length cDNA;

sequencing the obtained oligonucleotide ditag;

determining the ditag of interest; and

recovering the full-length cDNA corresponding to the ditag of interest from the full-length cDNA library.

The invention also provides a method for quantifying the transcriptional activity of a gene comprising:

preparing, from a full-length cDNA library, at least one oligonucleotide comprising at least one ditag, the ditag comprising two joined first and second sequence tags, wherein the first tag comprises the 5'-terminus sequence and the second tag comprises the 3'-terminus sequence of a full-length cDNA;

sequencing the obtained oligonucleotide ditag;

determining the frequency of the sequenced ditag which corresponds to the transcriptional activity of the gene.

Having now generally described the invention, the same will be more readily understood through reference to the following examples which are provided by way of illustration, and are not intended to be limiting of the present invention.

EXAMPLES

GIS oligonucleotides for cDNA synthesis, the structure of a generic 50 bp cohesive ditag, primers used for the construction of vector pGIS1, and ds-DNA adapter which are used in the examples are listed below.

GIS Analysis Oligos for cDNA synthesis

- GsuI-oligo dT primer:

5'-GAGCTCCTTCTGGAGTTTTTTTTTTTTTTTTVN-3' (SEQ ID NO: 1)

- NotI/BamHI/MmeI(N)6 primer linker (top):

5'-AATTCGCGGCCGCTTGGATCCGACNNNNNN (SEQ ID NO:2)

- NotI/BamHI/MmeI(N) primer linker (bottom):

5'-p-GTCGGATCCAAGCGGCCGCG-3' (SEQ ID NO:3)

- NotI/BamHI/MmeI(N)5 primer linker (top):

5'-AATTCGCGGCCGCTTGGATCCGACGNNNNN (SEQ ID NO:4)

- MmeI/BamHI/SalI adapter (top):

5'-TCGACCCAGGATCCAACTT-3' (SEQ ID NO:5)

- MmeI/BamHI/SalI adapter (bottom):

5'-p-GTTGGATCCTGGG – 3' (SEQ ID NO:6)

- PMR003: 5'-GTAAAACGACGGCCAGT-3' (SEQ ID NO:7)

- PMR004: 5'-GGAAACAGCTATGACCATG-3' (SEQ ID NO:8)

- PMR006: 5'-TAATACGACTCACTATAGGG-3' (SEQ ID NO:9)

- PMR011: 5'-GATGTGCTGCAAGGCGATTAAG-3' (SEQ ID NO: 10)

- PMR012: 5'-AGCGGATAACAATTTCACACAGG-3' (SEQ ID NO:11).

Structure of a generic 50bp cohesive ditag

5'-GATCCGACXXXXXXXXXXXXXXXXXXXNNNNNNNNNNNNNNNNAAGTTG

(SEQ ID NO:12)

GCTGXXXXXXXXXXXXXXXXXXXNNNNNNNNNNNNNNNNTTCAACCTAG -5'

(SEQ ID NO:13)

Wherein X and N may be any of A, C, G or T.

Primers used for the construction of vector pGIS1

Mme_mut1: 5'-p-CGCTCTCCTGTACCGACCCTGCCGCTTAC-3' (SEQ ID NO:14)

Mme_mut2: 5'-p-AACTATCGTCTTGAGACCAACCCGGTAAG-3' (SEQ ID NO:15)

ds-DNA adapter

5'-AATTCTCGAGCGGCCGCGATATCG-3' (SEQ ID NO:16)

3'-GAGCTCGCCGGCGCTATAGCTTAA-p-5' (SEQ ID NO:17)

pGIS1 sequence

The sequence of pGIS1 (SEQ ID NO:18) is shown in Figure 10.

EXAMPLE 1

The method

The experimental procedure of GIS ditag analysis has been carried out according to the following modules of cDNA library construction and analysis:

(1) The full-length cDNA library which introduces the MmeI sites flanking both ends of each cDNA insert;

(2) The GIS ditag library in which each clone contains a 5' 18bp signature and a 3' 18bp signature of a transcriptional unit;

(3) The GIS library for clones of concatenated GIS ditags;

(4) GIS sequencing analysis.

1. GIS full-length cDNA library with addition of MmeI sites for each cDNA inserts

The outline of procedure of this section was as follows: starting from high quality mRNA, the first cDNA was synthesized with a GsuI-oligo dT primer (SEQ ID NO:1).

The first strand cDNA/RNA hybrids was subjected to a full-length enrichment procedure by the biotinylation-based cap-trapper approach. Any cap-trapper approach known in the art can be used, for example Carninci et al., 1996, Genomics, Vol.37, 327-336; US 6,143,528; Edery et al., 1995, Mol. Cell. Biol., Vol.15, No.6, 3363-3371).

The enriched full-length first strand cDNA was the template for second cDNA synthesis primed with adapter-primer (NotI/BamHI/MmeI-(N)5 and -(N)6, (SEQ ID NOS: 2-4) that contain a MmeI, a BamHI, and a NotI site.

After the double strand cDNA was made, the cDNA poly-A/T tail was cleaved off by GsuI restriction enzyme. GsuI is another Type-II endonuclease that cleaves DNA 16bp from its recognition site. At the GsuI cleavage end, an adapter containing a MmeI, a BamHI site, and a SalI cohesive end was ligated to the cDNA (SEQ ID NOS: 5-7).

Following a NotI digestion, the full-length cDNA was inserted into the vector pGIS1, between the NotI and SalI sites in the polylinker. The vector pGIS1 (see Figures 7 and 10) is modified from pGEM3z (Promega).

1-1. mRNA preparation

The total mRNA has been prepared from mouse embyonic stem cell line E14 using Trizol reagent (Invitrogen). However, any standard method (as those described in Sambrook J. and Russell D.W., 2001, Molecular Cloning, Cold Spring Harbor Laboratory Press) may also be used.

mRNA (polyA RNA) was purified by oligo dT magnetic beads according to standard techniques (for example, Sambrook and Russell, 2001, as above). Alternatively, purification may be carried out by affinity column according to standard techniques (for example, Sambrook and Russell, 2001, as above).

1-2. First Strand cDNA synthesis and full-length selection

In this step, the first cDNA is synthesized with a GsuI-oligo dT primer. Then, the first strand cDNA/RNA hybrids are subjected to a full-length enrichment procedure by the biotinylation-based cap-trapper approach.

GsuI-oligo dT primer:

5'-GAGCTCCTTCTGGAGTTTTTTTTTTTTTTTTVN-3' (SEQ ID NO: 1)

The following were mixed:

GsuI-oligo dT primer (7 μg/μl)	2μl
PolyA RNA (20μg)	18μl

The obtained solution was heated to 65°C for 10min and 37°C for 1min.

Then, spin tube in microfuge and the following substances were added:

2X GC-I buffer (Takara)	75μl
RNase inhibitor Promega)	1μl
10mM dNTP (with methyl-dCTP)	4μl
Saturated trehalose	10μl
4.9M sorbitol	26μl
Superscript II reverse transcriptase (Invitrogen)	15ul

The obtained solution was incubated at 37°C for 10min, 42°C for 30min 50°C for 20 min and 55°C for 20min. 2 μl of proteinase K (20 mg/ml) were added. The obtained solution was Incubated at 45°C for 15 min followed by phenol/chloroform extraction and isopropanol precipitation (according to standard technique, eg. Sambrook and Russel, 2001, as above).

The RNA/cDNA heteroduplex was re-suspended into 44.5 μl of ddH_2O. 3 μl of 1.1 M NaOAc pH 4.5 and 2.5 μl of 100mM $NaIO_4$ were added to oxidize the diol structures of the mRNA. 50 μl of the reaction solution were incubated on ice in the dark for 45 min followed by adding 0.5 μl of 10% SDS, 11 μl of 5 M NaCl and 61 μL of isopropanol to precipitate the

RNA/DNA. The precipitated RNA/DNA was resuspended in 50 µl of ddH_2O. 5µL 1M NaOAc (pH6.1), 5µL 10% (w/v) SDS and 150µL 10mM long-arm biotin hydrazide were added to biotinylate the RNA. The reaction was incubated at room temperature in dark overnight. The biotinylated RNA/DNA was precipitated by adding 5µL 5M NaCl, 75µL 1M RNase-free NaOAc (pH6.1), and 750µL 100% EtOH or 200 µL of 100% Isopropanol. Incubate at -80°C for 30min by spin 14krpm at 4°C for 30min.

The pellet was washed w/ 70% (v/v) EtOH/30%, DEPC-treated ddH_20 (DEPC is diethylpyrocarbonate, which is an RNase inhibitor), and 14krpm spin was carried out at 4°C for 10min. The extra liquid was carefully removed. Then, the pellet was air-dried. and resuspended in 400µL DEPC- ddH_20. Then 50µL 10x RNaseI buffer and 25 units RNaseI / µg of starting mRNA were added. The obtained solution was incubated at 37°C for 30min. 10µL of 10mg/mL Yeast tRNA (Ambion) and 150µL of 5M NaCl were added to stop the reaction.

While biotinylating the RNA-DNA heteroduplex, the Streptavidin Dynabeads were prepared as follows: 400µL of M-280 Streptavidin beads (Dynal) were pipetted into an RNase-free Eppendorf tube, the beads placed on a magnet, left staying for at least 30min, and then the supernatant was removed. The beads were re-suspended in 400µL 1x binding buffer (2M NaCl, 50 mM EDTA, pH 8.0). The tube was placed on a magnet, waited at least 30min, and then the supernatant was removed. The 1x binding buffer wash was repeated for 2 more times. The beads were re-suspended in 400µL 1x binding buffer with 100µg of Yeast tRNA, and then incubated at 4°C for 30min with occasional mixing. The tube was placed on a magnet stand, waited at least 30 seconds, and the supernatant was removed. The beads were washed with 1x binding buffer for 3 times. The beads and RNA/DNA heteroduplex were mixed (the total volume now was 660µL, and the binding condition was at 1 M NaCl). The mixture was rotated at room temperature for 30min.

The tube was placed on a magnet stand, waited at least 30 seconds, and the supernatant removed (the supernatant was saved as "unbound").

The beads were washed two times with 400µL of 1x binding buffer. Washed with 400µL of 0.4%(w/v) SDS plus 50µg/mL Yeast tRNA. Washed with 400µL of 1x wash buffer (10mM Tris-HCl pH7.5, 0.2mM EDTA, 10mM NaCl & 20%(v/v) glycerol, 40 µg/mL Yeast tRNA).

And washed w/400μL of 50μg/mL Yeast tRNA. For all washes the tube was placed on a magnet stand, waited for at least 30 seconds, and the supernatant was removed.

The first strand cDNA was released by alkali hydrolysis of RNA. The following was added: 50μL 50mM NaOH and 5 mM EDTA (pH8.0). The tube was rotated at 65°C for 10min. The tube was placed on a magnet stand, and the supernatant transferred to another tube containing 50μL 1M Tris-Cl (pH7.5).

The lysis procedure was repeated for 2 more times. The final volume of supernatant was 300 μL (containing the first strand cDNA).

The cDNA was extracted by phenol/chloroform extraction and precipitate by 1mL ethanol with glycogen;. The cDNA was re-suspended in 5μL LoTE (0.1X) LoTE is low salt Tris-EDTA buffer (3mM Tris-HCl pH 7.5 and 0.2mM EDTA pH7.5)).

1-3. Second strand cDNA synthesis

The following reagents were added to the each corresponding tube on ice.

cDNA in LoTE	5μL
Linker (N5)	1.6μg
Linker (N6)	0.4μg
Soln II (Takara ligation kit)	10μL
Soln I (Takara ligation kit)	20μL

Linker (N6) is:

NotI/BamHI/MmeI(N)6 primer linker (top):

5'-AATTCGCGGCCGCTTGGATCCGACNNNNNN (SEQ ID NO:2)

NotI/BamHI/MmeI(N) primer linker (bottom):

5'-p-GTCGGATCCAAGCGGCCGCG-3' (SEQ ID NO:3)

Linker (N5) is:

NotI/BamHI/MmeI(N)5 primer linker (top):

5'-AATTCGCGGCCGCTTGGATCCGACGNNNNN (SEQ ID NO:4)

NotI/BamHI/MmeI(N) primer linker (bottom): is the sequence (SEQ ID NO:3) indicated above.

The cDNA and linker mixture was incubated at 16°C overnight. And the following were added:

ddH_20	20μL
10XExTaq buffer (Takara)	8μL
2.5mM dNTP	8μL
ExTaq polymerase (Takara)	4μL

The mixture was preheated in a thermo-cycler 65°C, 5min → 68°C, 30min → 72°C, 10min., followed by phenol/chloroform extraction and ethanol ppt with glycogen, and re-suspended in 85μl ddH_20.

1-4. Removal of polyA tail by GsuI digestion

The following reagents were added to the tube.

cDNA	85μL
GsuI (Fermentas)	5μL
10X bufferB (Fermentas)	10μL

The mixture was incubated at 30°C for 2 hours, followed by phenol/chloroform and ethanol precipitation. The pellet was re-suspended in 10ul ddH2O, and the following 3' adaptor ligation reaction was carried out.

1-5. Addition of 3' adaptor with MmeI and BamHI and SalI sites

The following components were added to the tube containing 10μl of sample. The 10μl of sample was the double-stranded full-length cDNA which has had the poly(A) tail removed by GsuI digestion.

5X ligation buffer	10μL
GsuI SalI adapter (0.4 μg /μL)	25μL

[The GsuI SalI adapter is MmeI/BamHI/SalI adapter)

T4 DNA ligase (5 units/ul) (Invitrogen)	5μL

- MmeI/BamHI/SalI adapter (top):

5'-TCGACCCAGGATCCAACTT-3' (SEQ ID NO:5)

- MmeI/BamHI/SalI adapter (bottom):

5'-GTTGGATCCTGGG-p-3' (SEQ ID NO:6)

The reaction was incubated at 16°C overnight, followed by phenol/chloroform extraction and ethanol precipitation, and the pellet re-suspended in 41 μl dH_2O.

1-6. NotI digestion and cDNA size fractionation

The following were added on ice and in order:

NEB Buffer 3	5μL
NotI (10units/μl) (NEB)	4μL

The obtained solution was incubated at 37°C for 1-2 hours.

cDNA Size Fractionation Columns were prepared (the Invitrogen instructions were followed: uncap the column (bottom first) and allow it to drain completely; wash 5 times with 800μL $T_{10}E_{0.1}N_{25}$ Buffer, allowing the column to drain completely each time). The DNA sample was loaded onto the column. The flow-through was collected in an Eppendorf tube (fraction 1). 100μL of $T_{10}E_{0.1}N_{25}$ Buffer were added. The flow-through was collected in an Eppendorf tube (fraction 2). Another 100μL of $T_{10}E_{0.1}N_{25}$ Buffer was added. The flow-through

collected, one drop per pre-numbered Eppendorf tube (beginning with fraction 3, each drop was about 30-40μL).

Whenever the column runs dry, another 100μL of $T_{10}E_{0.1}N_{25}$ Buffer may be added.

Up to drop 20 should be collected (according to the Invitrogen protocol). 3μL of each fraction were run on agarose gel to visualize the cDNA size in each fraction. Pool fractions were showing cDNA ≥1.0kbp (usually up to 2-3 kbp). Pooled samples were kept neat (using a cuvette soaked at least 30' in slightly acidified 100% EtOH, rinsed 5 times with ddH_2O, and saving sample. This is what has to be ligated).

If only one fraction is to be used, precipitate it and use the half to all of it, depending on what the gel looks like.

At this point the cDNA fragments have the NotI cohesive end at 5' side and SalI cohesive end at 3' side, and are ready to be cloned in vector.

1-7. Ligation of cDNA with linearized plasmid pGIS1.

1-7-1 The cloning vector pGIS1 was prepared by NotI and SalI digestion. The vector sequence of pGIS1 is shown in Figures 7 and 9.

<u>pGIS1 Cloning vector construction</u>

(I) Site-specific mutagenesis of pGEM3z to create MmeI-minus vector

The vector pGIS-1 was derived from pGEM3z (Promega). pGEM3z originally contained two MmeI recognition sites that were knocked-out by site-directed mutagenesis. The QuikChange Multi kit (Stratagene) was used, together with mutagenic primers:

Mme_mut1:

5'-p-CGCTCTCCTGT<u>ACCGAC</u>CCTGCCGCTTAC-3' (SEQ ID NO:14)

Mme_mut2:

5'-p-AACTATCgTCTTgAg<u>ACCAAC</u>CCggTAAg-3' (SEQ ID NO:15)

(II) Modification of polylinker region

The polylinker region was modified by simple insertion of a ds-DNA adapter at the existing EcoRI site. Additional recognition sites thus introduced are: XhoI, NotI and EcoRV (EcoRV is deleted upon insertion of the stuffer fragment (see below)).

ds-DNA adapter:

5'-AATTCTCGAGCGGCCGCGATATCG-3' (SEQ ID NO:16)

3'-GAGCTCGCCGGCGCTATAGCTTAA-p-5' (SEQ ID NO:17)

(III) Stuffer fragment insertion

An approximately 690bp fragment was inserted between the NotI and SalI sites of the modified vector (see vector sequence in Figure 10). This facilitated the production of NotI/SalI double-digested vector, as the stuffer can be clearly visualized and excised during gel-purification.

The linearized plasmid was gel purified.

1-7-2 The cDNA was ligated to the pGIS1 vector overnight and the constructs were transferred into electrocompetent *E. coli* TOP10 cells by electroporation according to standard techniques (see Sambrook and Russel, 2001, as above).

1-8. Library QC (QC = Quality Check)

A dilution series of 1-100µL of transformants was plated out onto LB agar plates with antibiotic selection. The colonies were incubated overnight and counted to determine the library titer.

Between 24 to 96 colonies (arbitrary numbers) were picked and the inserts size determined by direct colony PCR and agarose gel electrophoresis (according to standard techniques, eg. Sambrook and Russel, 2001, see above). The percentage of cDNA insert and the average insert size were estimated.

At this stage, the GIS full-length cDNA library may be stored as ligation reactions or as transformants in *E. coli* cells, according to standard methodology (Sambrook and Russel, 2001, see above).

EXAMPLE 2

2. GIS ditag library

The cDNA clones made from steps 1-1 to 1-8 contained a MmeI site (TCCGAC) at the 5' side and another MmeI site (TCCAAC) in reverse orientation at the 3' end. Note that these two MmeI recognition sites are two isoforms that can be recognized by MmeI (TCCRAC 20/18, where R = (A/G)). The sequence difference here will be useful later for directional indication. MmeI restriction enzyme will cleave these clones 20bp into the cDNA fragments from their 5' and 3' ends. Consequently, despite the variable sizes of the digested cDNA, the vector plus the 20bp cDNA signature tags on each end of all clones will be of a constant size that can be easily recognized upon agarose gel electrophoresis, and can be easily purified from the unwanted cDNA fragments.

The gel-purified vector plus tags can then be self-ligated to give a "tagged plasmid" containing the 5' and 3' GIS signature tags.

2-1. Plasmid preparation

The GIS full-length cDNA library was amplified once by plating an appropriate number of clones on large (22 x 22cm) agar plates (Genetix). The number of colonies required was determined by the estimated transcriptome size. After an overnight 37C incubation, the resultant bacterial colonies were harvested and pelleted by centrifugation at 3000g for 30min. Plasmid DNA preparation was performed using the Qiagen HiSpeed Plasmid Maxi kit. The quality of the DNA obtained was examined by agarose gel electrophoresis and restriction digestion. Approximately 300,000 colonies can be processed to yield at least 1mg of plasmid DNA.

2-2. MmeI digestion

Approximately 10μg of plasmid DNA was digested using MmeI as per manufacturer's conditions (NEB), ensuring that the number of units of enzyme used was always less than 4-fold excess to prevent methylation-induced inhibition. Digestion proceeded at 37C for 2-6 hrs.

An aliquot of the digestion reaction was examined on an agarose gel: a strong band of approximately 2800bp in size corresponding to the linearized vector containing the GIS

signature tags were easily observed, together with a number of fragments derived from the excision of cDNA from the original plasmids (see Figure 4).

2-3. Linear vector-GIS ditag purification

The digestion reaction was electrophoresed on 0.7% agarose, and the 2800bp vector-GIS tag band was excised and purified using the Qiagen agarose gel extraction kit.

2-4. Vector-GIS ditag self ligation to create "tagged-plasmids"

MmeI digestion resulted in a 2bp overhang on both the 5' and 3' signature tags. These were removed (polished off) using T4 DNA polymerase (Promega), leaving behind 18bp tags:

(0.5-2.0ug) DNA	50μL
10x Y+/TANGO buffer (Fermentas)	6.0μL
0.1M DTT	0.3μL
T4 DNA polymerase	5units/μg
10mM dNTP	0.6μL
ddH_2O	to 60.0μL

Incubated at 37C, for 5min, then inactivate at 75C for 10min

The purified, blunted DNA was then ethanol precipitated and resuspended at a concentration of approximately 20ng/μl. Self-ligation (intramolecular recircularization) was carried out as follows:

Approx. 350ng DNA	15.0μL
Ligation Solution I (Takara Ligation Kit 2)	15.0μL

Incubated at 16C, 2hr to overnight

2-5. Creation of Di-signature Tags (ditags)

The goal of this step was to obtain the GIS di-signature tags in a form quantitatively representative of the original cDNA library from which the tagged-plasmids were derived.

Structure of a generic 50bp cohesive ditag

5'-GATCCGACXXXXXXXXXXXXXXXXXXXNNNNNNNNNNNNNNNNAAGTTG

(SEQ ID NO:12)

GCTGXXXXXXXXXXXXXXXXXXXNNNNNNNNNNNNNNNNTTCAACCTAG -5'

(SEQ ID NO:13)

Wherein X and N may be any of A, C, G or T.

We used two approaches to this:

(i) Bacterial transformation, tagged-plasmid purification and restriction digest to release 50bp cohesive ditags;

(ii) Direct PCR on the ligation reaction followed by restriction digest of the PCR products to release 50bp cohesive ditags.

2-5-1 Transformation and propagation; Preparation of tagged-plasmids (See Figure 1)

1μl of the ligation reaction (Section 2-4) were transformed per 50μl of electrocompetent TOP10 cells (Invitrogen) by electroporation. Recovered in 1ml SOC media at 37C for 1hr, then plated out several dilutions on LB agar + ampicillin for QC and titering.

QC (Quality Check) : plasmid DNA was prepared from several colonies and tested by digestion with BamHI: tagged-plasmids release a 50bp cohesive ditag.

This process was then scaled-up by plating the remaining culture on large agar plates, and performing maxipreps using Qiagen HiSpeed Plasmid Maxi kit.

As an example, approximately 5,000 colonies was processed to yield at least 40ug of tagged-plasmid DNA.

This plasmid DNA was then BamHI-digested to generate 50bp cohesive ditags (see Figure 5 as example result).

2-5-2 PCR-based retrieval of cohesive ditags (See Figure 2

PCR was performed on the ligation reaction using primers PMR003 and PMR004 that bind to vector sequences flanking the ditags.

PMR003: 5'-GTAAAACGACGGCCAGT-3' (SEQ ID NO:7)

PMR004: 5'-GGAAACAGCTATGACCATG-3' (SEQ ID NO:8)

The amount of starting material was determined empirically by doing a series of dilutions and choosing the conditions that result in a clean, specific PCR product of approximately 200bp

(e.g. 1:200) diluted ligation reaction 5.0μL

10x HiFi buffer	2.0μL
10mM dNTP	0.4μL
PMR003 (100ng/μL)	1.0μL
PMR004 (100ng/μL)	1.0μL
Eppendorf TripleMaster polymerase	0.2μL
dH2O	10.4μL

(the HiFi buffer was the reaction buffer provided with the Eppendorf TripleMaster enzyme)

Thermo-cycling conditions:

Step1: 95C x 2min

Step 2: 95C x 30sec

Step 3: 55C x 1min

Step 4: 72C x 30sec

Go to step 2, repeat steps (2-4) 24x

Step 5: 72C x 4min

16^0C forever

The PCR products were analyzed on a 1.5% agarose gel.

For negative controls, the PCR reaction was performed using (i) no template, and (ii) no ligase. To obtain sufficient 200bp PCR product for subsequent 50bp cohesive ditag production, the reaction was scaled-up: do 96 PCR reactions using a 96-well PCR plate; this generates approx. 50ug of 200bp ditag. The individual PCR reactions are then combined and ethanol precipitated before Bam HI digest to generate 50bp cohesive ditags (see figure 6 as an example result).

3. GIS library

3-1. Tagged-plasmid preparation

This applies only to the bacterial transformation-based approach (see Section 2-5-1).

3-2. BamHI digestion and purification of GIS tags

3-2-1 BamHI digestion of tagged-plasmids (Section 2-5-1) released 50bp cohesive ditags:

DNA (tagged-plasmids)	40μg
10x unique BamHI buffer (NEB)	100μL
100x BSA	10μL
BamHI (20U/μL, NEB)	10μL
dH2O	to 1mL

The choice of value of 40μg of DNA (tagged-plasmids) was arbitrary.

Aliquots were divided into 10x 100ul for more efficient digestion, and incubated at 37C, for 4hrs.

After digest, they were inactivated at 65C, for 15min, then phenol-chloroform extraction and ethanol precipitation were performed. Then, the pellet comprising 50bp cohesive ditags and the rest of the cleavage products after the BamHI digest was resuspended in LoTE buffer for gel-purification.

3-2-2 BamHI digestion of or 200bp ditags retrieved by PCR (Section 2-5-2) released 50bp cohesive ditags:

DNA (PCR products)	40μg
10x unique BamHI buffer (NEB)	100μL
100x BSA	10μL
BamHI (20U/μL, NEB)	10μL
dH2O	to 1mL

The choice of value of 40μg of DNA (tagged-plasmids) was arbitrary.

Aliquots were divide into 10x 100ul for more efficient digestion, incubated at 37C, for 4hrs.

After digest, they were inactivated at 65C, for 15min, then phenol-chloroform extraction and ethanol precipitation were performed. Then, the pellet comprising 50bp cohesive ditags and the rest of the cleavage products after the BamHI digest was resuspended in LoTE buffer for gel-purification.

3-3 Gel-purification of 50bp cohesive ditags

The BamHI-digested DNA according to both section 3-2-1 or 3-2-2 was separated on a large (Hoefer Ruby 600, 15 x 15cm, 1.5mm thick) 10% polyacrylamide gel. Electrophoresis proceeded at 200V for approx. 2hrs until the Bromophenol Blue (standard tracking dye) band almost reached the bottom of the gel. The gel was stained in SYBR Green I (Molecular Probes, Inc.) for 30min before visualisation and excision of the 50bp cohesive ditags.

At this stage it is convenient not to load more than 5μg per lane, or ~~the~~ fluorescence quenching occurs.

The 50bp cohesive ditags were excised and collected into 0.6ml microfuge tubes (2 gel pieces per tube) which have been pierced at the bottom with a 21G needle. This pierced tube was placed inside a 1.7ml microfuge tube, and centrifuged at 12K g, 4C for 2-5min. The gel pieces were thus shredded and collected in the 1.7ml tube.

150µl of LoTE:NH_4OAc (125:25) were added to each tube and left overnight at 4C to elute. The next day, the eluate was collected with the aid of microspin filter units (SpinX, Costar), and ethanol precipitation performed to retrieve the purified 50bp ditags, which were resuspended in LoTE. Starting from 70µg 200bp ditag, we expected to retrieve several hundred ng of 50bp ditag.

3-4. Ditag concatenation and gel-purification

Some optimization (ligation time, amount of starting material) may be necessary to ensure that the concatenation of the 50bp ditags results in a smear of products ranging from approx. 300bp to >1000bp. The conditions below are suggested as a starting point:

50bp cohesive ditags	150- 500ng
5x buffer (with PEG; BRL)	2.0µL
T4 DNA ligase (5U/µL)	1.0µL
dH2O	to 10µL

Incubated at 16^0C for 1hr.

Loading buffer was added and the entire sample heated at 65C for 15min. The sample loaded in a single well of an 8% polyacrylamide minigel and run at 200V for about 1 hr, or until Bromophenol Blue was about 2 cm from bottom.

The smear of ligation products can be excised as 2 or more fractions, eg. 200-500 bp; 500-1000bp; >1000bp.

Elution of DNA from the gel pieces was performed as detailed in Section 3-3. The eluate was extracted with phenol-chloroform then ethanol precipitated. Resuspend the DNA pellet in 6ul LoTE.

3-5. Cloning of concatemers

The cloning vector was prepared by digesting 2ug of pZErO-1 plasmid DNA (Invitrogen) (Figure 8) (Figure 8 shows the sequencing/ PCR primer binding sites) with 10 units of BamHI for 3 hours at 37C. The digested DNA was phenol-chloroform extracted and ethanol

precipitated, then resuspended in LoTE at a concentration of 33ng/μl. The ligation reaction was performed as follows:

Concatemer DNA	6.0μL
BamHI/pZErO-1	1.0μL
5x ligase buffer	2.0μL
T4 DNA ligase (5U/μL)	1.0μL

Incubated at 16^0C overnight.

The vector self-ligation was also performed in parallel as a control.

The ligation products were purified before electroporation. The phenol-chloroform extraction was followed by ethanol precipitation; the pellet was washed 3 times with 75% ethanol before re-suspending in 12μl TE (0.1X). 1μl of this DNA was used to transform 50μl of electro-competent TOP10 bacterial cells. After recovery (see also Section 2-5-1), 50μl were plated on a small agar plate (containing Low Salt LB agar (Lennox L) plus Zeocin (50μg/ml) and IPTG (50μg/ml) and incubated overnight at 37C. As a background control, bacteria were plated out that have been similarly transformed with the vector self-ligation reaction above. (IPTG is optional when using TOP10 cells but may reduce background).

3-6. GIS library QC (Quality Check)

The following day, 10-30 colonies were picked to check for insert size by PCR. For each reaction, a single colony was picked into a PCR tube containing:

10x HiFi buffer	2.0μL
10mM dNTP	0.4μL
PMR003 (100ng/ul)	1.0μL
PMR004 (100ng/ul)	1.0μL
Eppendorf TripleMaster polymerase	0.2μL
dH2O	11.4μL

Thermo-cycling conditions:

Step1: 95C x 2min

Step 2: 95C x 30sec

Step 3: 55C x 1min

Step 4: 72C x 3min

Go to step 2, repeat steps (2-4) 24x

Step 5: 72C x 4min

16^{0}C forever

The PCR products were visualized on 1% agarose gel. A typical result is shown in Figure 9.

The primer pair PMR003/PMR004 (SEQ ID NO:7/SEQ ID NO:8) gives a band of approx. 220bp in the absence of any cloned insert. If the quality of the library thus produced appears good, the remaining transformation mixture can be plated out (Section 3-5) on large agar plates in preparation for DNA sequencing analysis.

The primer pair PMR003/PMR004 is also convenient for checking the quality of the library, but for the actual preparation of PCR templates for sequencing, primer pair PMR012/PMR003 (SEQ ID NO:11/SEQ ID NO:7) were preferred (see Section 4-2).

PMR012: 5'-AGCGGATAACAATTTCACACAGG-3' (SEQ ID NO:11).

4. Sequencing analysis of GIS tags

4-1. Library plating and colony picking

The transformed TOP10 (Invitrogen) bacteria cells were plated out on 22x22cm agar plates with colony density less than 3,000 per plate. Individual colonies were picked and cultured in 384-well plates with LB plus Zeocin (see above in section 3.5) at 37^{0}C overnight. Multiple copies of 384-well plates are replicated and stored in -80^{0}C.

4-2. Template preparation

Bacterial cultures in 384-well plates were inoculated in pre-mixed PCR cocktails. PCR was performed using primer pair PMR012/PMR003.

This primer pair gives a band of 245bp in the absence of any concatemer insert. Nonetheless, this set of primers is preferred as it allows the use of sequencing primers PMR004 (M13 reverse; 68bp from BamHI site) and PMR006 (SEQ ID NO:9)(M13 forward; 87bp from BamHI site).

PMR006: 5'-TAATACGACTCACTATAGGG-3' (SEQ ID NO:9)

4-3. Sequencing

PCR templates were sequenced using the sequencing primers PMR004 and PMR006 to sequence in both directions.

EXAMPLE 3

The GIS analysis method according to any embodiment of the invention is a complete gene discovery platform. It combines full-length cDNA library construction, cDNA tag sequencing, genome mapping and annotation into one operation from the same starting materials. For example, to study the genes expressed in human stem cells, we start with the stem cell mRNA, construct a stem cell GIS full-length cDNA library, and then the GIS library. We will only need to sequence 50,000 clones of the GIS library to reveal over a million transcripts. Such deep sampling will allow us to capture nearly all unique transcripts expressed in the human stem cell transcriptome. Each of the GIS ditags can be specifically mapped to the genome and therefore define the structural regions of the corresponding genes on the chromosomes. Most of the GIS ditags map to known genes on chromosomes and the counts of the GIS ditags provide the measurement of expression activity. Some of the GIS ditags may map to desert ("no gene") regions of the genome, which may suggest the identification of new genes that are expressed in the stem cell transcriptome. In such a way the genome annotation for genes is further refined by this whole transcriptome-to-whole genome approach. Based on the GIS ditag sequences, these putative new genes can be readily cloned from the original GIS full-length cDNA library.

We can apply this GIS gene discovery system not only to human stem cells, but also to all other biological systems, such as development of cells, tissues and organs of human and model organisms.